

(19)



(11)

EP 3 929 758 B1

(12)

EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention of the grant of the patent:

23.04.2025 Bulletin 2025/17

(51) International Patent Classification (IPC):

G06F 13/12 ^(2006.01) **H04L 9/40** ^(2022.01)
H04L 49/10 ^(2022.01) **H03K 19/17728** ^(2020.01)
H04L 47/10 ^(2022.01) **H04L 49/90** ^(2022.01)

(21) Application number: **20210908.8**

(52) Cooperative Patent Classification (CPC):

H04L 49/10; G06F 13/128; H04L 63/1458;
H03K 19/17728; H04L 47/10; H04L 49/90;
H04L 63/0428

(22) Date of filing: **01.12.2020**

(54) **STACKED DIE NETWORK INTERFACE CONTROLLER CIRCUITRY**

NETZWERKSCHNITTSTELLENSTEUERUNGSSCHALTUNG FÜR GESTAPELTE CHIPS

CIRCUIT DE COMMANDE D'INTERFACE DE RÉSEAU DE PUCE EMPILÉE

(84) Designated Contracting States:

AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO
PL PT RO RS SE SI SK SM TR

- **Ravipalli, Sreedhar**
Santa Clara, CA 95054 (US)
- **Cheerla, Rakesh**
Santa Clara, CA 95054 (US)
- **Horne, Martin**
Santa Clara, CA 95054 (US)

(30) Priority: **26.06.2020 US 202016914164**

(43) Date of publication of application:

29.12.2021 Bulletin 2021/52

(74) Representative: **Goddar, Heinz J.**

Boehmert & Boehmert
Anwaltpartnerschaft mbB
Pettenkoferstrasse 22
80336 München (DE)

(73) Proprietor: **INTEL Corporation**

Santa Clara, CA 95054 (US)

(72) Inventors:

- **Zaman, Naveed**
Santa Clara, CA 95054 (US)
- **Dasu, Aravind**
Santa Clara, CA 95054 (US)

(56) References cited:

US-A1- 2019 043 536 US-A1- 2019 044 519
US-A1- 2019 230 049

Note: Within nine months of the publication of the mention of the grant of the European patent in the European Patent Bulletin, any person may give notice to the European Patent Office of opposition to that patent, in accordance with the Implementing Regulations. Notice of opposition shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

Description

Background

[0001] This relates to integrated circuits, and more particularly, to integrated circuits used to implement network interface cards.

[0002] Conventional network interface cards are typically implemented either as an application-specific integrated circuit (ASIC) die or as a field-programmable gate array (FPGA) die. The ASIC implementation provides performance, power, and area advantages but its customized nature for a specific use is oftentimes not flexible enough to support a wide range of different customer requirements. The FPGA implementation provides more configurability to meet different customer requirements but at the cost of decreased performance, more power consumption, and increased area/cost.

US 2019 / 0230049 A1 discloses integrated circuit devices that may include a network processor in a data processing die and an on-package memory in a base die. The data processing die may implement one or more network functionalities that may exchange data with low-latency memory, high capacity in the base die. The data processing die may be programmable fabric, which may be dynamically reconfigured during operation. The document discloses that the data processing die is stacked vertically with respect to the base die and that the base die includes basic network function circuits. However, the document does not disclose that the base die is an application-specific integrated circuit (ASIC) die comprising non-customizable network functions.

[0003] The document further does not disclose a flow rate control hardware accelerator circuit configured to drop packets that are taking up more than a predetermined chunk of bits on a network port. Furthermore, dropping or ignoring packets is not disclosed, as is the use of network ports.

[0004] It is within this context that the embodiments described herein arise.

Summary

[0005] The invention is set out in the appended set of claims.

Brief Description of the Drawings

[0006]

FIG. 1 is a diagram of an illustrative system of integrated circuit devices operable to communicate with one another in accordance with an embodiment. FIG. 2 is a diagram of an illustrative system with a network interface controller (NIC) that includes customizable networking blocks and primitive functional blocks in accordance with an embodiment. FIG. 3 is a cross-sectional side view of an illustrative

network interface controller (NIC) implemented in a stacked die configuration in accordance with an embodiment.

FIG. 4 is a diagram of an illustrative programmable integrated circuit die that includes customizable networking blocks in accordance with an embodiment. FIG. 5 is a diagram of an illustrative application-specific integrated circuit (ASIC) die that includes primitive functional blocks in accordance with an embodiment.

FIG. 6 is a diagram of an illustrative programmable integrated circuit in accordance with an embodiment. FIG. 7 is a top plan (layout) view of the ASIC die shown in FIG. 5 in accordance with an embodiment. FIGS. 8A and 8B show illustrative steps for operating a network interface controller of the type shown in connection with FIGS. 2-7 in accordance with an embodiment.

Detailed Description

[0007] The present embodiments relate to a network interface controller (NIC) implemented using a programmable integrated circuit die stacked on top of an application-specific integrated circuit (ASIC) die. This stacked die configuration combines the benefits of both ASICs and programmable logic devices (PLDs) by using the programmable interconnect fabric in the top programmable die to implement customer-specific requirements and forming static fixed functional blocks on the bottom ASIC die (e.g., a highly flexible packet processing pipeline can be implemented using the programmable fabric of the top die, whereas stable, high area, and high power consumption blocks can be implemented as hardened blocks on the bottom die without burdening the general purpose programmable fabric on the top die).

[0008] By stacking the programmable die directly on top of the ASIC die, close coupling between the two dies can be achieved, which enables efficient inter-block control and data communication between the two dies. The bottom (base) die may also include an extensive network of memory circuits that serve as lookup tables and intermediate packet buffers. Network interface controllers implemented in this way can provide superior customer configurability/customizability while also meeting power and form factor constraints

[0009] FIG. 1 is a diagram of an illustrative system 100 of interconnected electronic devices. The system of interconnected electronic devices may have multiple electronic devices such as device A, device B, device C, device D, and interconnection resources 102. Interconnection resources 102 such as conductive lines and busses, optical interconnect infrastructure, or wired and wireless networks with optional intermediate switching circuitry may be used to send signals from one electronic device to another electronic device or to broadcast information from one electronic device to multiple other electronic devices. For example, a transmitter in

device B may transmit data signals to a receiver in device C. Similarly, device C may use a transmitter to transmit data to a receiver in device B.

[0010] The electronic devices may be any suitable type of electronic device that communicates with other electronic devices. Examples of such electronic devices include integrated circuits having electronic components and circuits such as analog circuits, digital circuits, mixed-signal circuits, circuits formed within a single package, circuits housed within different packages, circuits that are interconnected on a printed-circuit board (PCB), etc.

[0011] To facilitate communications between an integrated circuit and a computing network, a network interface controller (sometimes referred to as a network interface card or "NIC") is provided. Typical smart-NIC processing of packets received from the network involves pre-parsing Ethernet packets, validating the packets, extracting quality of service (QoS) requirements from the packets, and decrypting the packets. The decrypted packets are stored in a buffer and the associated header information is passed to a packet processing engine. The packet processing engine transforms the packet to enable network address and port translation services, connection tracking, policing, and gathering of per-entity statistics.

[0012] Smart-NICs implemented using only a programmable integrated circuit such as a field-programmable gate array (FPGA) die provides the desired flexibility to enable different open-virtual switching and virtual router implementations as well as being configurable to optimize for the network service. Network interface cards are also used to facilitate secure communication channels between different compute and storage servers by accelerating or offloading cryptographic functions, remote direct memory access, flow scheduling, and compression services. As the packet processing rate increases, however, these basic packet processing functions tend to consume a lot of FPGA resources, which translate to higher logic element usage and higher power consumption.

[0013] In accordance with an embodiment, a network interface controller (NIC) is provided where the basic primitive functions are implemented on a base (bottom) die i.e. an application-specific integrated circuit or "ASIC" die) while the configurable fabric and packet processing components are implemented on a top (stacked) die (i.e., a programmable integrated circuit such as a field-programmable gate array or "FPGA"). The programmable fabric of the top die provides high bandwidth interconnect circuitry to efficiently access the primitive functions on the base die.

[0014] FIG. 2 is a diagram of an illustrative system 200 with a network interface controller such as NIC 208. As shown in FIG. 2, network interface controller 208 may be interposed between a host processor such as host processor 202 and a computing network such as network 220. In general, system 200 may be part of a personal area network (PAN), a local area network (LAN), a me-

tropolitan area network (MAN), a wide area network (WAN), a storage area network (SAN), an enterprise private network (EPN), a datacenter network, a virtual private network (VPN), a system area network, or other suitable types of data storage and processing networks. Host processor 202 may for example be a central processing unit (CPU), a microprocessor, a microcontroller, or a graphics processing unit (GPU). Host processor 202 (sometimes referred to as a host CPU or simply CPU) may include one or more processing cores for processing instructions of a computer program or software application.

[0015] Host processor 202 may include a host direct memory access (DMA) component such as host DMA engine 204 for direct access of an external host memory 206. Host memory 206 is sometimes referred to as a "main memory" device. Main memory refers to physical memory that can be directly accessed by host processor 202. Main memory 206 may be implemented using volatile memory components such as dynamic random-access memory (DRAM). Main memory (sometimes also referred to as "primary storage") is distinguished from external mass storage devices such as disk drives, optical drives, and tape drives. CPU 202 can only manipulate data that is stored in main memory 206. Thus, every program that is executed or every file that is accessed must be copied from an external mass storage device into main memory 206. The amount of storage in memory main 206 may therefore determine how many programs can be executed at any point in time and the amount of data that can be made readily available to the program.

[0016] In general, the software running on host CPU 202 may be implemented using software code stored on non-transitory computer readable storage media (e.g., tangible computer readable storage media). The software code may sometimes be referred to as software, data, program instructions, instructions, script, or code. The non-transitory computer readable storage media may include non-volatile memory such as non-volatile random-access memory (NVRAM), one or more hard drives (e.g., magnetic drives or solid state drives), one or more removable flash drives or other removable media, or the like. Software stored on the non-transitory computer readable storage media may be executed on the processing circuitry of host processor 202.

[0017] Still referring to FIG. 2, network interface controller 208 may include a host interface such as host interface 208 operable to communicate with host processor 202, a memory interface such as memory interface 216 operable to communicate with off-chip NIC-attached memory devices 218, a network interface such as network interface(s) 214 operable to communicate with network 220. Network interface controller 208 includes customizable networking circuits 210, and primitive functional circuits 212.

[0018] Host interface 210 may be a non-coherent computer bus interface such as the PCIe (Peripheral Component Interconnect Express) interface or a coherent

computer bus interface such as UltraPath Interconnect (UPI), QuickPath Interconnect (QPI), Compute Express Link (CXL), Cache Coherent Interconnect for Accelerators (CCIX), Gen-Z, Open Coherent Accelerator Processor Interface (OpenCAPI), Intel Accelerator Link (IAL), NVidia's NVLink, or other computer bus interfaces.

[0019] Memory interface 216 may be implemented as double data rate (DDR) interfaces such as DDR type-3 (DDR3), low power DDR3 (LPDDR3), DDR type-4 (DDR4), low power DDR4 (LPDDR4), DDR type-5 (DDR5), graphics DDRx, quad data rate (QDR), Open NAND Flash Interface (ONFI), or other suitable interfaces for communicating with external memory. Memory interface 216 may include memory controllers for supporting a wide variety of external memory protocols. Network interface(s) 214 may be implemented as a wired IEEE 802.3 Ethernet interface, a Fast Ethernet interface, a Gigabit Ethernet interface, a Token Ring interface (just to name a few), or a wireless connection (e.g., a IEEE 802.11 Wi-Fi or Bluetooth® connection).

[0020] The customizable networking circuits 210 include circuit blocks that might evolve from one product generation to another and/or that might be different depending on a customer's needs. In contrast, the "primitive" function circuits 212 represent basic/generic networking blocks that remain stable between product generations and/or remain fixed across different customer applications. As such, the customizable (and evolving) networking circuit blocks 210 are implemented using "soft" or reconfigurable resources on a programmable integrated circuit die such as a field-programmable gate array (FPGA), whereas the primitive (and stable) functional circuit blocks 212 is implemented as "hardened" or nonreconfigurable functional blocks on an application-specific integrated circuit (ASIC) die.

[0021] As demands on integrated circuit technology continue to outstrip even the gains afforded by ever decreasing device dimensions, an increasing number of applications demand a packaged solution with more integration than is possible in one silicon die. In an effort to meet this need, more than one IC die may be placed within an integrated circuit package (i.e., a multichip package). As different types of devices cater to different types of applications, more IC dies may be required in some systems to meet the requirements of high performance applications. Accordingly, to obtain better performance and higher density, a multichip package may include multiple dies arranged laterally along the same plane; the present invention, the multichip package includes multiple dies stacked on top of one another.

[0022] FIG. 3 is a cross-sectional side view of an illustrative network interface controller (NIC) such as NIC 208 of FIG. 2 implemented as a stacked multichip package such as multichip package 300. As shown in FIG. 3, multichip package 300 includes a top FPGA die 302 and associated external memory 218 stacked directly on top of a bottom ASIC die 304 (sometimes referred to as the "base die"). This configuration in which

top die 302 is stacked vertically over base die 304 is sometimes referred to as a 3-dimensional or "3D" die stack configuration.

[0023] The customizable (and evolving) networking circuit blocks 210 are implemented on the top die 302, whereas the primitive (and stable/fixed) functional circuit blocks 212 are implemented on the bottom die 304. This close coupling between top die 302 and bottom die 304 facilitates efficient inter-block control and data communication between the two dies. The primitive hardened function blocks on the base die 304 may be selectively accessed using flexible high-bandwidth interconnect fabric on the top die 302. This arrangement in which the high power and high logic element consuming functions are allocated to the base die while the flexible interconnect circuitry are delegated to the top FPGA die creates a customizable packet processing pipeline while meeting desirable power and form factor targets.

[0024] The arrangement of FIG. 3 in which the programmable die 302 is stacked on top of the ASIC die 304 is merely illustrative. If desired, the ASIC die 304 may alternatively be stacked on top of the programmable die 302.

[0025] FIG. 4 is a diagram showing illustrative circuit blocks that may be included within the top programmable die 302. As shown in FIG. 4, top programmable die 302 may include a host interface and associated direct memory access (DMA) engine (labeled collectively as circuits 400 and optionally corresponding to at least host interface 210 of FIG. 2), network interface(s) 402 (optionally corresponding to network interface 214 of FIG. 2), memory interfaces 404 (optionally corresponding to memory interface 216 of FIG. 2), packet buffering and arbitration hardware acceleration circuits 406, flexible packet processing pipeline and high bandwidth interconnects 408, customer application intellectual property (IP) blocks 410, and other suitable configurable network processing and storage circuitry.

[0026] Packet buffering and arbitration hardware (HW) acceleration (ACC) circuits 406 may include ingress arbiter and class of service (CoS) queue(s), collectively labeled as circuits 407-A. Circuits 406 may also include egress arbiter and CoS queue(s), collectively labeled as circuits 407-B. Ingress arbiter and CoS queue circuits 407-A may be primarily responsible for absorbing short term high packet incoming rate and preventing important data packets from dropping when packets cannot be processed at a lower rate due to resource constraints. Circuits 407-A may send flow rate control messages to the remote sender to ensure that no packets of such class of packets are getting dropped. Egress arbiter and CoS queue circuits 407-B may be primarily responsible for handling receipt of remote flow control messages and stalling transmission of packets in response to receive such flow control messages. Circuits 407-B may also be responsible for prioritizing the scheduling of important packets onto the network relative to traffic with lesser priority.

[0027] The flexible (reconfigurable) packet processing pipeline and high bandwidth interconnects 408 may serve as a programmable yet efficient conduit between the different functional blocks on the base die. If desired, the packet processing pipeline of the top die may optionally be configured to support packet parsing, packet validation, extracting flow keys for exact matching operations, longest prefix matching (LPM) operations, and range and wildcard match based lookup to extract actions on how to modify/remove/add bytes of a packet, connection tracking, policing and statistics gathering operations, flexible congestion control for remote DMA, and other suitable operations that might be desirable to be implemented using programmable logic resources. Customer application IP block 410 may be configured to support whatever application that is desired by the user of the overall NIC component.

[0028] The top die 302 may be implemented as a programmable integrated circuit such as a programmable integrated circuit 10 of the type shown in FIG. 6. Programmable integrated circuit 10 may be a programmable logic device (PLD). Examples of programmable logic devices include programmable arrays logic (PALs), programmable logic arrays (PLAs), field programmable logic arrays (FPLAs), electrically programmable logic devices (EPLDs), electrically erasable programmable logic devices (EEPROMs), logic cell arrays (LCAs), complex programmable logic devices (CPLDs), and field programmable gate arrays (FPGAs), just to name a few. As shown in FIG. 6, device 10 may include a two-dimensional array of functional blocks, including logic array blocks (LABs) 11 and other functional blocks, such as random access memory (RAM) blocks 13 and specialized processing blocks such as digital signal processing (DSP) blocks 12 that are partly or fully hardwired to perform one or more specific tasks such as mathematical/arithmetic operations.

[0029] Functional blocks such as LABs 11 may include smaller programmable regions (e.g., logic elements, configurable logic blocks, or adaptive logic modules) that receive input signals and perform custom functions on the input signals to produce output signals. Device 10 may further include programmable routing fabric that is used to interconnect LABs 11 with RAM blocks 13 and DSP blocks 12. The combination of the programmable logic and routing fabric is sometimes referred to as "soft" logic, whereas the DSP blocks are sometimes referred to as "hard" logic. The type of hard logic on device 10 is not limited to DSP blocks and may include other types of hard logic. Adders/subtractors, multipliers, dot product computation circuits, and other arithmetic circuits which may or may not be formed as part of a DSP block 12 may sometimes be referred to collectively as "arithmetic logic."

[0030] Programmable logic device 10 may contain programmable memory elements for configuring the soft logic. Memory elements may be loaded with configuration data (also called programming data) using input/output

elements (IOEs) 16. Once loaded, the memory elements provide corresponding static control signals that control the operation of one or more LABs 11, programmable routing fabric, and optionally DSPs 12 or RAMs 13. In a typical scenario, the outputs of the loaded memory elements are applied to the gates of metal-oxide-semiconductor transistors (e.g., pass transistors) to turn certain transistors on or off and thereby configure the logic in the functional block including the routing paths. Programmable logic circuit elements that may be controlled in this way include parts of multiplexers (e.g., multiplexers used for forming routing paths in interconnect circuits), look-up tables, logic arrays, AND, OR, NAND, and NOR logic gates, pass gates, etc. The logic gates and multiplexers that are part of the soft logic, configurable state machines, or any general logic component not having a single dedicated purpose on device 10 may be referred to collectively as "random logic."

[0031] The memory elements may use any suitable volatile and/or non-volatile memory structures such as random-access-memory (RAM) cells, fuses, antifuses, programmable read-only-memory memory cells, mask-programmed and laserprogrammed structures, mechanical memory devices (e.g., including localized mechanical resonators), mechanically operated RAM (MORAM), programmable metallization cells (PMCs), conductive-bridging RAM (CBRAM), resistive memory elements, combinations of these structures, etc. Because the memory elements are loaded with configuration data during programming, the memory elements are sometimes referred to as configuration memory, configuration RAM (CRAM), configuration memory elements, or programmable memory elements.

[0032] In addition, programmable logic device 10 may use input/output elements (IOEs) 16 to drive signals off of device 10 and to receive signals from other devices. Input/output elements 16 may include parallel input/output circuitry, serial data transceiver circuitry, differential receiver and transmitter circuitry, or other circuitry used to connect one integrated circuit to another integrated circuit. As shown, input/output elements 16 may be located around the periphery of the chip. If desired, the programmable logic device may have input/output elements 16 arranged in different ways.

[0033] The routing fabric (sometimes referred to as programmable interconnect circuitry) on PLD 10 may be provided in the form of vertical routing channels 14 (i.e., interconnects formed along a vertical axis of PLD 10) and horizontal routing channels 15 (i.e., interconnects formed along a horizontal axis of PLD 10), each routing channel including at least one track to route at least one wire. If desired, the functional blocks of device 10 may be arranged in more levels or layers in which multiple functional blocks are interconnected to form still larger blocks. Other device arrangements may use functional blocks that are not arranged in rows and columns.

[0034] FIG. 5 is a diagram showing illustrative circuit blocks that are included within the base ASIC die 304.

[0035] As shown in FIG. 5, bottom die 304 includes a flow rate control hardware accelerator circuit 508 and base memory circuits 500 and may further include primitive functional blocks including but not limited to: transportation encryption/decryption hardware accelerator circuit 502, storage compression/decompression (at-rest encryption) hardware accelerator circuit 504, network flow search hardware accelerator circuit 506, distributed denial of service (DDoS) hardware accelerator circuit 510, and other suitable high power or stable/fixed functional blocks that are better suited for an ASIC implementation. If desired, in an alternative example not covered by the claim, the bottom die 304 may also be provided with one or more configurable logic sector blocks to offer more flexibility.

[0036] Base memory circuits 500 may be volatile memory circuit blocks such as static random-access memory (SRAM) blocks, buffers, cache, and/or other memory circuit suitable for storing packet information.

[0037] Transportation encryption/decryption (security) hardware accelerator 502 may help secure in-transmit networking application traffic between applications/containers across different hosts connected through an Ethernet network with protocols like Internet Protocol Security (IPsec), Datagram Transport Layer Security (DTLS), Advanced Encryption Standard (AES), Secure Hash Algorithms (SHA), and other SHA-like algorithms. As an example, transportation encryption/decryption HW accelerator 502 may include an inline cryptographic engine (ICE) that accelerates security protocol processing by implementing IPsec or security functions for DTLS. For instance, the inline crypto engine may implement 256-bit AES-GCM (Galois/Counter Mode), AES-CBC (Cipher Block Chaining), or SM4-like encryption algorithms to create a secure infrastructure. The inline crypto engine may also implement message authentication and anti-reply checking to prevent "man-in-the-middle" attacks.

[0038] Storage compression/decompression hardware accelerator 504 may compress and/or decompress network traffic, which helps reduce the size of stored data and therefore reduces the total cost of ownership for data center operators. Large data files can optionally be secured using "at-reset" encryption. As an example, storage compression/decompression HW accelerator 504 may include a remote DMA (RDMA) engine that enables reduction in CPU cycles in disaggregated storage applications. As another example, HW accelerator 504 may also include a look-aside cryptographic/compression engine (LCE) that enables compression and other cryptographic functions.

[0039] Network flow search hardware accelerator 506 may be used to search packet fields. A network flow is typically characterized by N different unique packet fields. Oftentimes, N-type packet fields can be hundreds or even thousands of bits. Network flow search HW accelerator 506 may perform such search using complex hashing algorithms. As an example, network flow search

HW accelerator 506 may include a cryptographic transport parser (CXP) that identifies a security association corresponding to a particular application packet. The CXP may extract different packet fields that are unique for an application and then perform a flow search operation to identify the security key(s) associated with the application. As another example, accelerator 506 may include a small exact match (SEM) engine or a large exact match (LEM) engine that is used to search a specific flow to extract flow keys from incoming packets. As yet another example, accelerator 506 may include a longest prefix match (LPM) engine that performs a longest prefix match for Internet Protocol version 4 (IPv4) and Internet Protocol version 6 (IPv6) addresses (as examples) to identify the final destination of a packet.

[0040] Flow rate control hardware accelerator 508 is used to ensure smooth delivery of service for all tenants on the network by controlling the traffic flow of an "aggressive" tenant. In other words, some networking application flows corresponding to one tenant or a group of tenants over-utilizing the network may be rate-controlled or throttled to ensure service level agreement (SLA) guarantees are met for all tenants hosted on the same host processor. As an example, accelerator 508 may include a traffic shaper engine that is used to regular the amount of traffic on a per application basis to guarantee a smoother delivery as well as reducing packet drops in the network. As another example, accelerator 508 may include a meters/policing and statistics engine that regulates the application traffic to enable SLA guarantees to other containers and applications while enabling gathering of accounting, billing, and other desirable networking analytics.

[0041] Distributed denial of service (DDoS) hardware accelerator 510 may be used to block service attacks originating from unauthorized network end points. As an example, DDoS HW accelerator 510 may include a range match (RM) engine and/or a wild card match (WCM) engine that block inappropriate application packets from suspicious applications to prevent potential DDoS attacks.

[0042] If desired, the base ASIC die 304 may also include accelerator circuits for performing connection tracking, key extraction, and other potentially high power or high logic resource consuming operation that might be more suitably offloaded to the base die.

[0043] FIG. 7 is a top plan (layout) view of bottom ASIC die 304. As shown in FIG. 7, ASIC die 304 has a variety of primitive network function blocks which may be distributed throughout an array of memory circuits 500 (see shaded blocks). For example, a transportation encryption/decryption hardware acceleration circuit block 502 (e.g., an inline cryptography engine) may be formed in a first location on die 304 surrounded by neighboring memory blocks 500, a first storage compression/decompression hardware accelerator circuit block 504-1 (e.g., an RDMA engine) may be formed in a second location on die 304 surrounded by neighboring memory blocks 500, a

second storage compression/decompression hardware accelerator circuit block 504-2 (e.g., a look-aside cryptography/compression engine) may be formed in a third location on die 304 surrounded by neighboring memory blocks 500, a network flow search hardware accelerator circuit block 506 (e.g., a crypto transport parser, an SEM/LEM engine, or LPM engine) and DDoS hardware accelerator circuit block 510 may both be formed in a fourth location on die 304 surrounded by neighboring memory blocks 500, a first flow rate control hardware accelerator circuit block 508-1 (e.g., a traffic shaper engine) may be formed in a fifth location on die 304 surrounded by neighboring memory blocks 500, and a second flow rate control hardware accelerator circuit block 508-2 (e.g., a RM/WCM engine) may be formed in a sixth location on die 304 surrounded by neighboring memory blocks 500.

[0044] The example of FIG. 7 in which six hardened primitive network function blocks are formed in a sea of base memory blocks 500 is merely illustrative. If desired, each non-memory function block need not be completely surrounded by memory blocks 500 (e.g., a given primitive block might only be surrounded by memory blocks 500 on three sides, on two sides, or on only one side), non-memory function blocks can be formed at any desired interval on die 304, any number of accelerator blocks can be grouped at any particular location, etc. If desired, the base die may also be implemented as multiple chiplets or smaller integrated circuit dies stacked below the top FPGA die. Each chiplet may include one or more of the primitive functional blocks and/or base memory circuits.

[0045] FIGS. 8A and 8B show illustrative steps for operating a network interface controller of the type shown in connection with FIGS. 2-7 in accordance with an embodiment. At step 800, a packet is received at network interface 214 from the network, which can then be fed to ingress arbiter and CoS queue circuits 407-A in the top die. Circuits 407-A may be primarily response for absorbing short term high packet incoming rate and preventing important data packets from dropping when packets cannot be processed at a lower rate due to resource constraints. Circuits 407-A may send flow rate control messages to the remote sender to ensure that no packets of such class of packets are getting dropped.

[0046] At step 802, the packet may be pre-classified to identify the desired class of service (e.g., by extracting CoS bits from the packet header or label field). At step 804, the packet may optionally be dropped or a transmit rate control message may be send to a remote transmitter to avoid drops.

[0047] At step 806, the NIC may arbitrate across network interface ports and CoS queues to select a next packet to be processed. The selected packet may then be send to a security protocol processing engine, which may include a cryptographic transport parser (CXP) and an inline cryptographic engine (ICE) as an example. In general, the security protocol processing engine may

include circuits from transportation encryption/decryption HW accelerator 502, network flow search HW accelerator 506, storage compression/decompression HW accelerator 504, and/or DDoS HW accelerator 510. At step 808, the security protocol processing engine may process the packet to extract security flow specific fields (e.g., to extract one or more security keys from the packet).

[0048] At step 810, the inline cryptographic engine (ICE) may use the extracted security key to decrypt and authenticate the packet. The ICE may help accelerate security protocol processing by implementing IP-sec or security functions for DTLS. For example, the inline crypto engine may implement 256-bit AES-GCM (Galois/Counter Mode), AES-CBC (Cipher Block Chaining), or SM4-like encryption algorithms to create a secure infrastructure. The inline crypto engine may also implement message authentication and anti-reply checking to prevent "man-in-the-middle" attacks.

[0049] At step 812, flow matching operations may be performed on the decrypted packet. As examples, different fields are matched against security or denial-of-service policies using the range matching (RM) engine and/or the wild card matching (WCM) engine. The RM/WCM engines may block inappropriate application packets from suspicious applications to prevent potential DDoS attacks. At step 814, the packet is optionally dropped from further processing if a security policy is violated.

[0050] Additional steps are shown in FIG. 8B. At step 816, if a security policy is not violated, the packet is sent to network flow search hardware accelerator 506 to determine where to send the packet. For example, the extracted flow keys may be used to perform exact or longest prefix matches to extract actions on how to process the packet (e.g., matching operations may be performed using the small exact match engine, the large exact match engine, and/or the longest prefix match engine).

[0051] At step 818, actions are extracted from the flow matching operations and metering is performed to prevent a particular flow from taking up an excessive amount of bandwidth. Packets that are taking up more than a predetermined chunk of bits on a network port are selectively dropped. The metering is managed using flow rate control hardware accelerator 508 (e.g., a traffic shaper engine or a meters/policing and statistics engine), which is used to ensure smooth delivery of service for all tenants on the network by controlling the traffic flow of an aggressive tenant. In other words, some networking application flows corresponding to one tenant or a group of tenants over-utilizing the network may be rate-controlled or throttled to ensure service level agreement (SLA) guarantees are met for all tenants hosted on the same host processor.

[0052] At step 820, the packet may be classified using the extracted actions, and the packet may either be sent to a storage processing flow/thread (proceeding to step 822) or to a non-storage processing flow/thread (pro-

ceeding to step 828).

[0053] When forking to the storage processing thread, the packet may be sent from the packet processor to an RDMA engine assuming an RDMA enabled flow (step 822). The packet processor may include circuits in only the top die (e.g., the flexible pipeline), in only the bottom die (e.g., the SEM/LEM engine, LPM engine, meters and statistics engine, etc.), or in both the top and bottom dies. At step 824, the packet may be sent from the RDMA engine to a storage processor. The storage processor may include circuits in only the top die, in only the bottom die, or in both the top and bottom dies. At step 826, the packet may be sent from the storage processor to application/software (SW) buffers (e.g., SW buffers within the customer IP blocks in the top die or application buffers in the host processor).

[0054] When forking to the non-storage processing thread, the packet may be sent from the packet processor to the DMA engine (e.g., a DMA engine on the top die or a DMA engine in the host processor). At step 830, the packet may be sent from the DMA engine directly to application buffers (e.g., application buffers within the customer IP blocks in the top die or software buffers in the host processor). For instance, the DMA engine may fetch descriptors to move the packet into the SW/application buffers and may send an interrupt notification to the application.

[0055] Follow either step 826 or 830, processing may proceed to step 832. At step 832, the application software may process the application data and may then release the application buffer back to the descriptor pool. Operated in this way, future packets may use the released application buffer pointed by such descriptor.

[0056] Although the methods of operations are described in a specific order, it should be understood that other operations may be performed in between described operations, described operations may be adjusted so that they occur at slightly different times or described operations may be distributed in a system which allows occurrence of the processing operations at various intervals associated with the processing, as long as the processing of the overlay operations are performed in a desired way.

[0057] For instance, all optional features of the apparatus described above may also be implemented with respect to the method or process described herein.

Claims

1. A network interface controller (208), comprising:

a first integrated circuit die (302) that includes customizable networking circuits (210), wherein the first integrated circuit die comprises a programmable integrated circuit die; and
a second integrated circuit die (304) that includes basic non-customizable network (220)

function circuits (212) wherein the second integrated circuit die is an application-specific integrated circuit, ASIC, die, wherein the basic non-customizable network function circuits in the second integrated circuit die comprises memory circuits (500) and a flow rate control hardware accelerator circuit (508) configured to drop packets that are taking up more than a predetermined chunk of bits on a network port, and wherein the first integrated circuit die is stacked vertically with respect to the second integrated circuit die.

2. The network interface controller (208) of claim 1, wherein the first integrated circuit die is stacked directly on top of the second integrated circuit die.

3. The network interface controller (208) of claim 1, wherein the second integrated circuit die is stacked directly on top of the first integrated circuit die.

4. The network interface controller (208) of any one of claims 1-3, wherein the first integrated circuit die comprises:

a host interface (400) operable to communicate with an external host processor;
a network interface (402) operable to communicate with a network (220); and
a memory interface (404) operable to communicate with an external memory device.

5. The network interface controller (208) of any one of claims 1-4, wherein the customizable networking circuits (210) in the first integrated circuit die comprises:
packet buffering and arbitration hardware acceleration circuits (406).

6. The network interface controller (208) of any one of claims 1-5, wherein the customizable networking circuits (210) in the first integrated circuit die comprises:
flexible packet processing pipeline and interconnect circuitry (408).

7. The network interface controller (208) of any one of claims 1-6, wherein the customizable networking circuits (210) in the first integrated circuit die comprises:
a customer application intellectual property (IP) block (410).

8. The network interface controller (208) of any one of claims 1 to 7, wherein the basic non-customizable network function circuits in the second integrated circuit die further comprises:
a transportation encryption and decryption hardware

accelerator circuit (502).

9. The network interface controller (208) of any one of claims 1-8, wherein the basic non-customizable network function circuits in the second integrated circuit die further comprises:
a storage compression and decompression hardware accelerator circuit (504). 5
10. The network interface controller (208) of any one of claims 1-9, wherein the basic non-customizable network function circuits in the second integrated circuit die further comprises:
a network (220) flow search hardware accelerator circuit (506). 10
11. The network interface controller (208) of any one of claims 1-10, wherein the basic non-customizable network function circuits in the second integrated circuit die further comprises:
a distributed denial of service hardware accelerator circuit (510). 20

Patentansprüche 25

1. Netzwerkschnittstellensteuerung (208), umfassend:
einen ersten integrierten Schaltung-Die (302), der anpassbare Vernetzungsschaltungen (210) beinhaltet, wobei der erste integrierte Schaltung-Die einen programmierbaren integrierten Schaltung-Die umfasst; und einen zweiten integrierte Schaltung-Die (304), der grundlegende nicht anpassbare Netzwerk- (220) Funktionsschaltungen (212) beinhaltet, wobei der zweite integrierte Schaltung-Die ein anwendungsspezifischer integrierter Schaltung-Die, ASIC-Die, ist, wobei die grundlegenden nicht anpassbaren Netzwerkfunktionsschaltungen in dem zweiten integrierten Schaltung-Die Speicherschaltungen (500) und eine Flussratensteuerungs-Hardware-Beschleunigerschaltung (508) umfassen, die dazu ausgelegt sind, Pakete, die mehr als einen vorbestimmten Bitblock auf einem Netzwerkport aufnehmen, und wobei der erste integrierte Schaltung-Die vertikal mit Bezug auf den zweiten integrierten Schaltung-Die gestapelt ist. 30
2. Netzwerkschnittstellensteuerung (208) nach Anspruch 1, wobei der erste integrierte Schaltung-Die direkt auf dem zweiten integrierten Schaltung-Die gestapelt ist. 50
3. Netzwerkschnittstellensteuerung (208) nach Anspruch 1, wobei der zweite integrierte Schaltung-Die direkt auf dem ersten integrierten Schaltung-Die gestapelt ist. 55
4. Netzwerkschnittstellensteuerung (208) nach einem

der Ansprüche 1-3, wobei der erste integrierte Schaltung-Die Folgendes umfasst:

eine Host-Schnittstelle (400), die betreibbar ist, um mit einem externen Host-Prozessor zu kommunizieren;
eine Netzwerkschnittstelle (402), die betreibbar ist, um mit einem Netzwerk (220) zu kommunizieren; und
eine Speicherschnittstelle (404), die betreibbar ist, um mit einer externen Speichervorrichtung zu kommunizieren.

5. Netzwerkschnittstellensteuerung (208) nach einem der Ansprüche 1-4, wobei die anpassbaren Netzwerkschaltungen (210) in dem ersten integrierten Schaltung-Die Folgendes umfassen:
Paketpuffer- und Arbitrierungs-Hardware-Beschleunigungsschaltungen (406). 15
6. Netzwerkschnittstellensteuerung (208) nach einem der Ansprüche 1-5, wobei die anpassbaren Netzwerkschaltungen (210) in dem ersten integrierten Schaltung-Die Folgendes umfassen:
flexible Paketverarbeitungspipeline und Verbindungsschaltungsanordnung (408). 25
7. Netzwerkschnittstellensteuerung (208) nach einem der Ansprüche 1-6, wobei die anpassbaren Netzwerkschaltungen (210) in dem ersten integrierten Schaltung-Die Folgendes umfassen:
einen Kundenanwendungsblock für geistiges Eigentum (IP) (410). 30
8. Netzwerkschnittstellensteuerung (208) nach einem der Ansprüche 1-7, wobei die grundlegenden nicht anpassbaren Netzwerkfunktionsschaltungen in dem zweiten integrierten Schaltung-Die ferner Folgendes umfassen:
eine Transportverschlüsselungs- und Entschlüsselungs-Hardware-Beschleunigerschaltung (502). 40
9. Netzwerkschnittstellensteuerung (208) nach einem der Ansprüche 1-8, wobei die grundlegenden nicht anpassbaren Netzwerkfunktionsschaltungen in dem zweiten integrierten Schaltung-Die ferner Folgendes umfassen:
eine Speicherkomprimierungs- und Dekomprimierungs-Hardware-Beschleunigerschaltung (504). 45
10. Netzwerkschnittstellensteuerung (208) nach einem der Ansprüche 1-9, wobei die grundlegenden nicht anpassbaren Netzwerkfunktionsschaltungen in dem zweiten integrierten Schaltung-Die ferner Folgendes umfassen:
eine Netzwerk- (220) Flusssuche-Hardware-Beschleunigerschaltung (506). 55

11. Netzwerkschnittstellensteuerung (208) nach einem der Ansprüche 1-10, wobei die grundlegenden nicht anpassbaren Netzwerkfunktionsschaltungen in dem zweiten integrierten Schaltung-Die ferner Folgendes umfassen:
eine verteilte Denial-of-Service-Hardware-Beschleunigerschaltung (510).

Revendications

1. Contrôleur d'interface réseau (208), comprenant :

une première puce de circuit intégré (302) qui comprend des circuits de réseau personnalisables (210), où la première puce de circuit intégré comprend une puce de circuit intégré programmable ; et

une seconde puce de circuit intégré (304) qui comprend des circuits de fonction de réseau de base non personnalisables (220) (212), où la seconde puce de circuit intégré est une puce de circuit intégré spécifique à une application, ASIC, où les circuits de fonction de réseau de base non personnalisables dans la seconde puce de circuit intégré comprennent des circuits de mémoire (500) et un circuit accélérateur matériel de contrôle de débit (508) configuré pour abandonner les paquets qui occupent plus qu'un bloc de bits prédéterminé sur un port réseau, et où la première puce de circuit intégré est empilée verticalement par rapport à la seconde puce de circuit intégré.

2. Contrôleur d'interface réseau (208) selon la revendication 1, dans lequel la première puce de circuit intégré est empilée directement sur la seconde puce de circuit intégré.

3. Contrôleur d'interface réseau (208) selon la revendication 1, dans lequel la seconde puce de circuit intégré est empilée directement sur la première puce de circuit intégré.

4. Contrôleur d'interface réseau (208) selon l'une quelconque des revendications 1 à 3, dans lequel la première puce de circuit intégré comprend :

une interface hôte (400) pouvant être utilisée pour communiquer avec un processeur hôte externe ;
une interface réseau (402) pouvant être utilisée pour communiquer avec un réseau (220) ; et
une interface mémoire (404) pouvant être utilisée pour communiquer avec un dispositif de mémoire externe.

5. Contrôleur d'interface réseau (208) selon l'une quel-

conque des revendications 1 à 4, dans lequel les circuits de mise en réseau personnalisables (210) dans la première puce de circuit intégré comprennent :

des circuits d'accélération matérielle de mise en mémoire tampon et d'arbitrage de paquets (406).

6. Contrôleur d'interface réseau (208) selon l'une quelconque des revendications 1 à 5, dans lequel les circuits de mise en réseau personnalisables (210) dans la première puce de circuit intégré comprennent :

un pipeline de traitement de paquets flexible et des circuits d'interconnexion (408).

7. Contrôleur d'interface réseau (208) selon l'une quelconque des revendications 1 à 6, dans lequel les circuits de mise en réseau personnalisables (210) dans la première puce de circuit intégré comprennent :

un bloc de propriété intellectuelle (IP) d'application client (410).

8. Contrôleur d'interface réseau (208) selon l'une quelconque des revendications 1 à 7, dans lequel les circuits de fonction réseau de base non personnalisables dans la seconde puce de circuit intégré comprennent en outre :

un circuit accélérateur matériel de chiffrement et de déchiffrement de transport (502).

9. Contrôleur d'interface réseau (208) selon l'une quelconque des revendications 1 à 8, dans lequel les circuits de fonction réseau de base non personnalisables dans la seconde puce de circuit intégré comprennent en outre :

un circuit accélérateur matériel de compression et de décompression de stockage (504).

10. Contrôleur d'interface réseau (208) selon l'une quelconque des revendications 1 à 9, dans lequel les circuits de fonction réseau de base non personnalisables dans la seconde puce de circuit intégré comprennent en outre :

un circuit accélérateur matériel de recherche de flux réseau (220) (506).

11. Contrôleur d'interface réseau (208) selon l'une quelconque des revendications 1 à 10, dans lequel les circuits de fonction réseau de base non personnalisables dans la seconde puce de circuit intégré comprennent en outre :

un circuit accélérateur matériel de déni de service distribué (510).

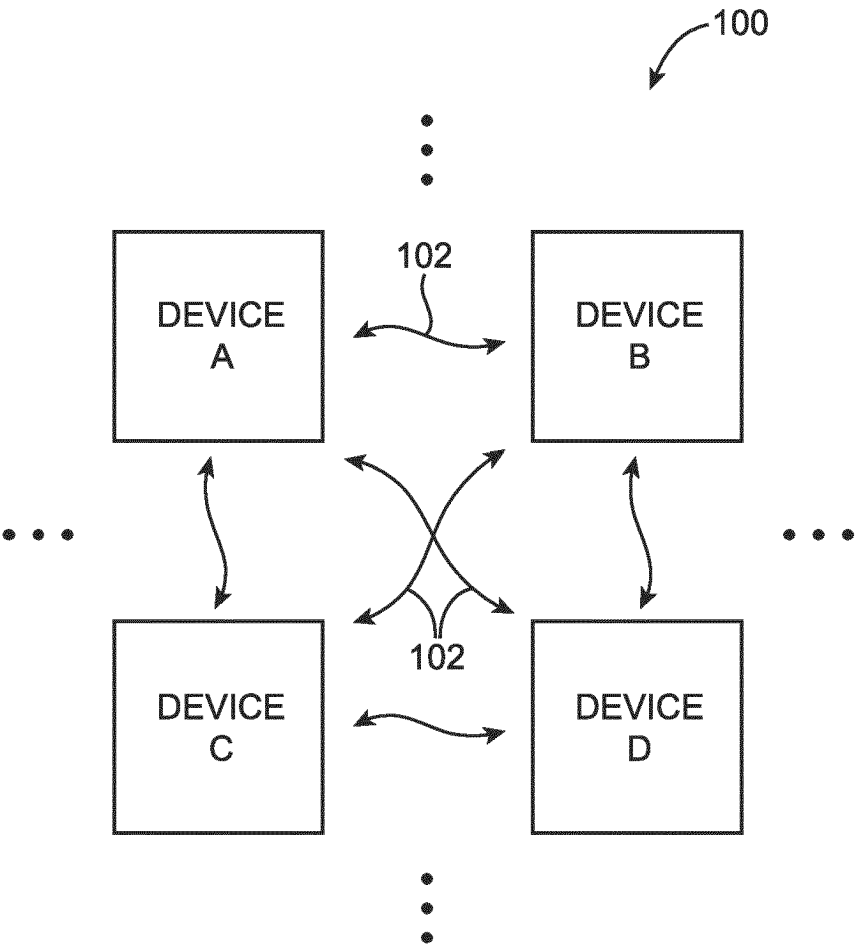


FIG. 1

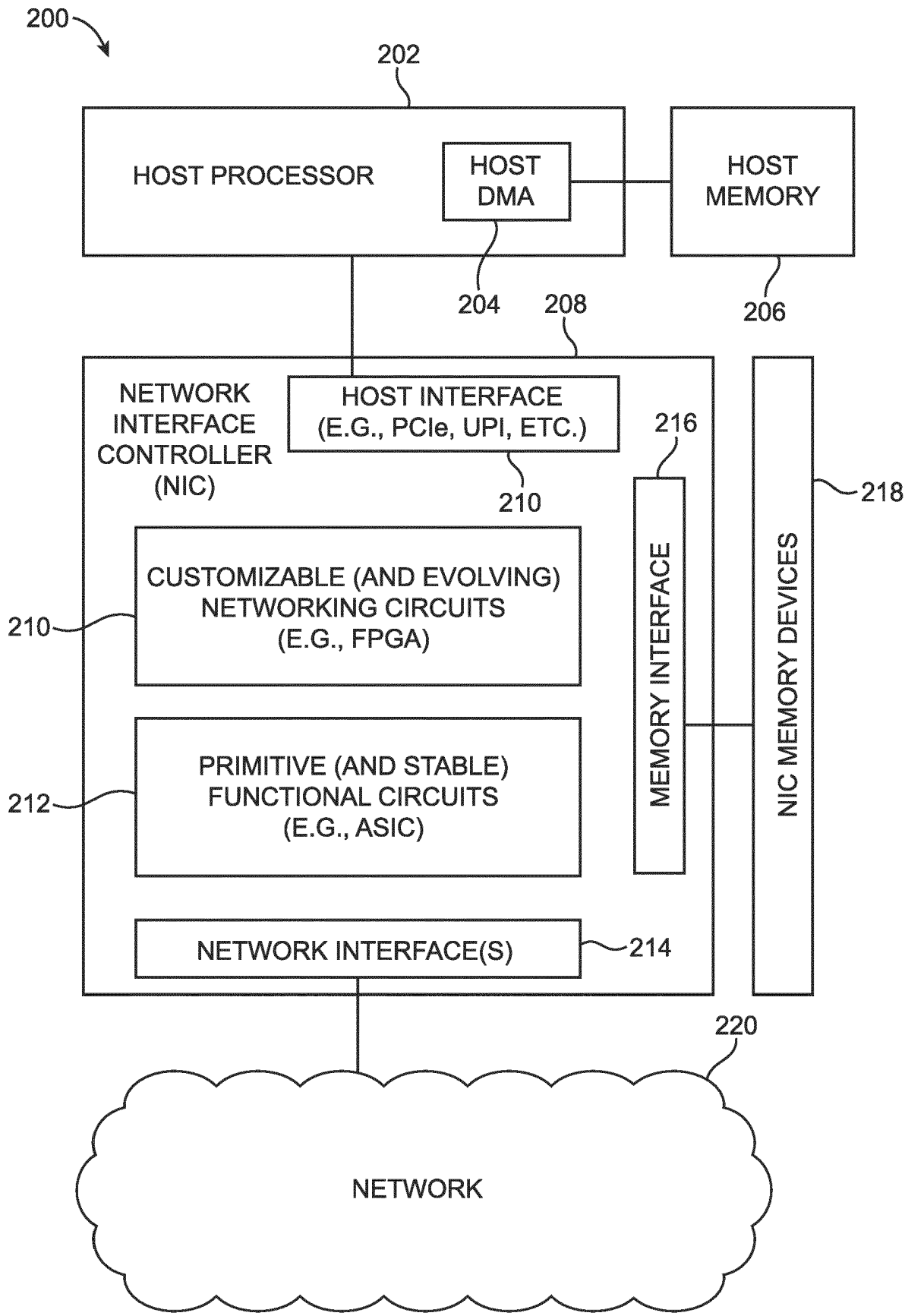


FIG. 2

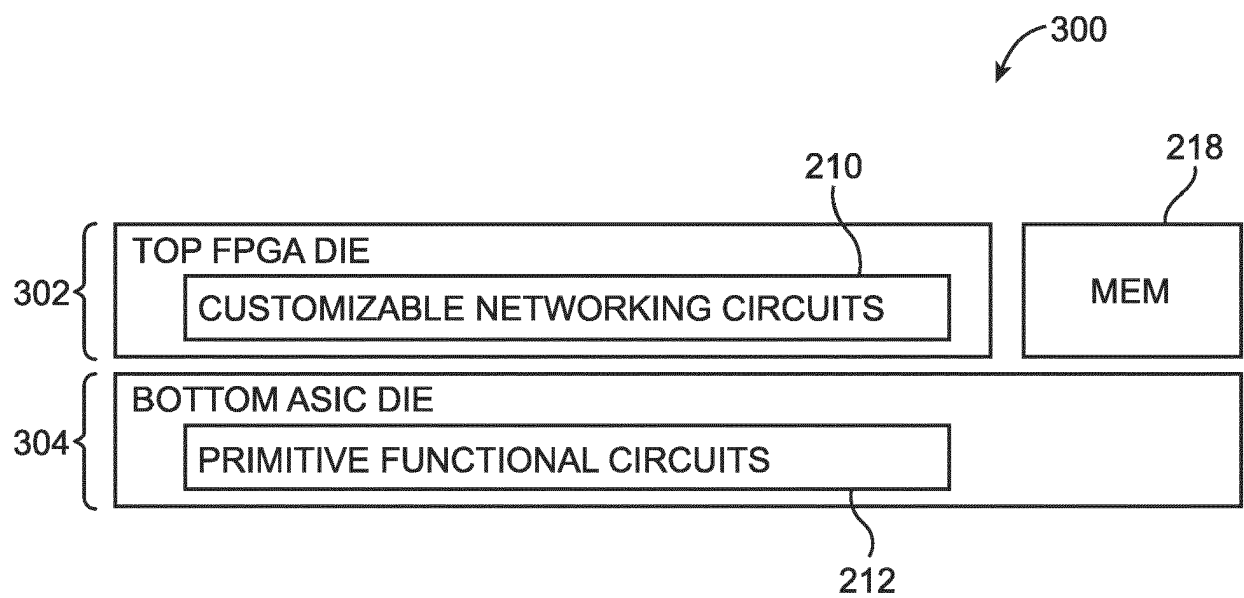


FIG. 3

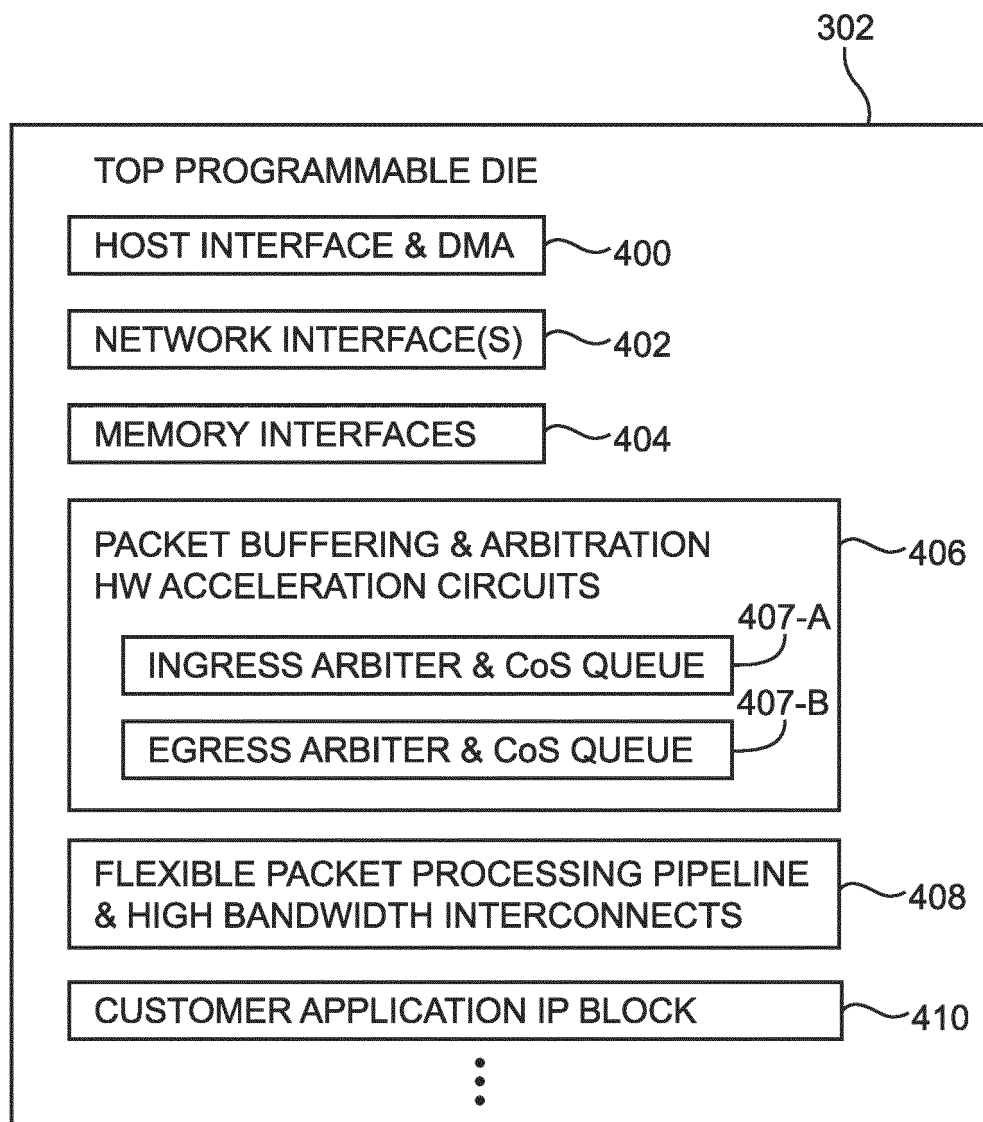


FIG. 4

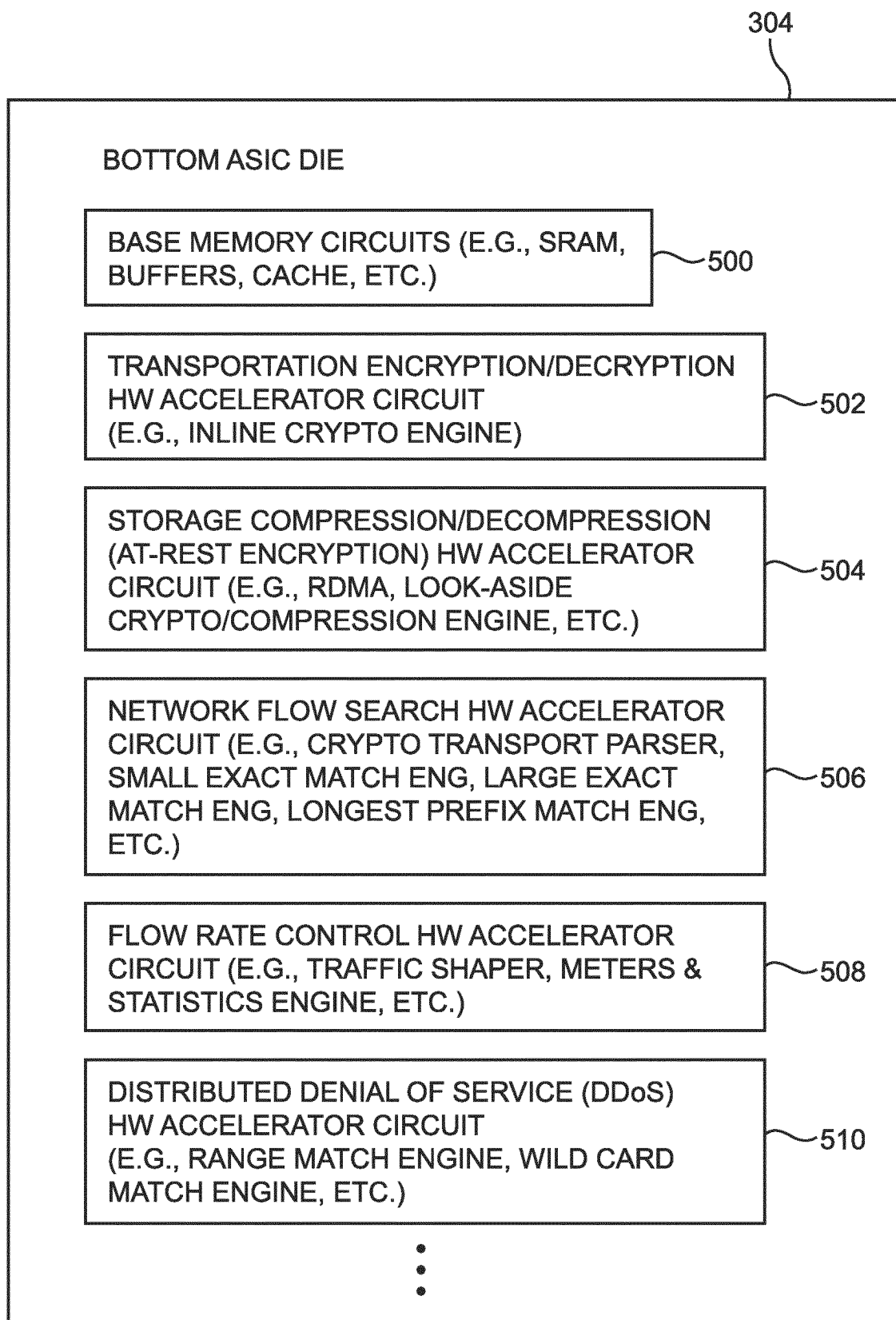


FIG. 5

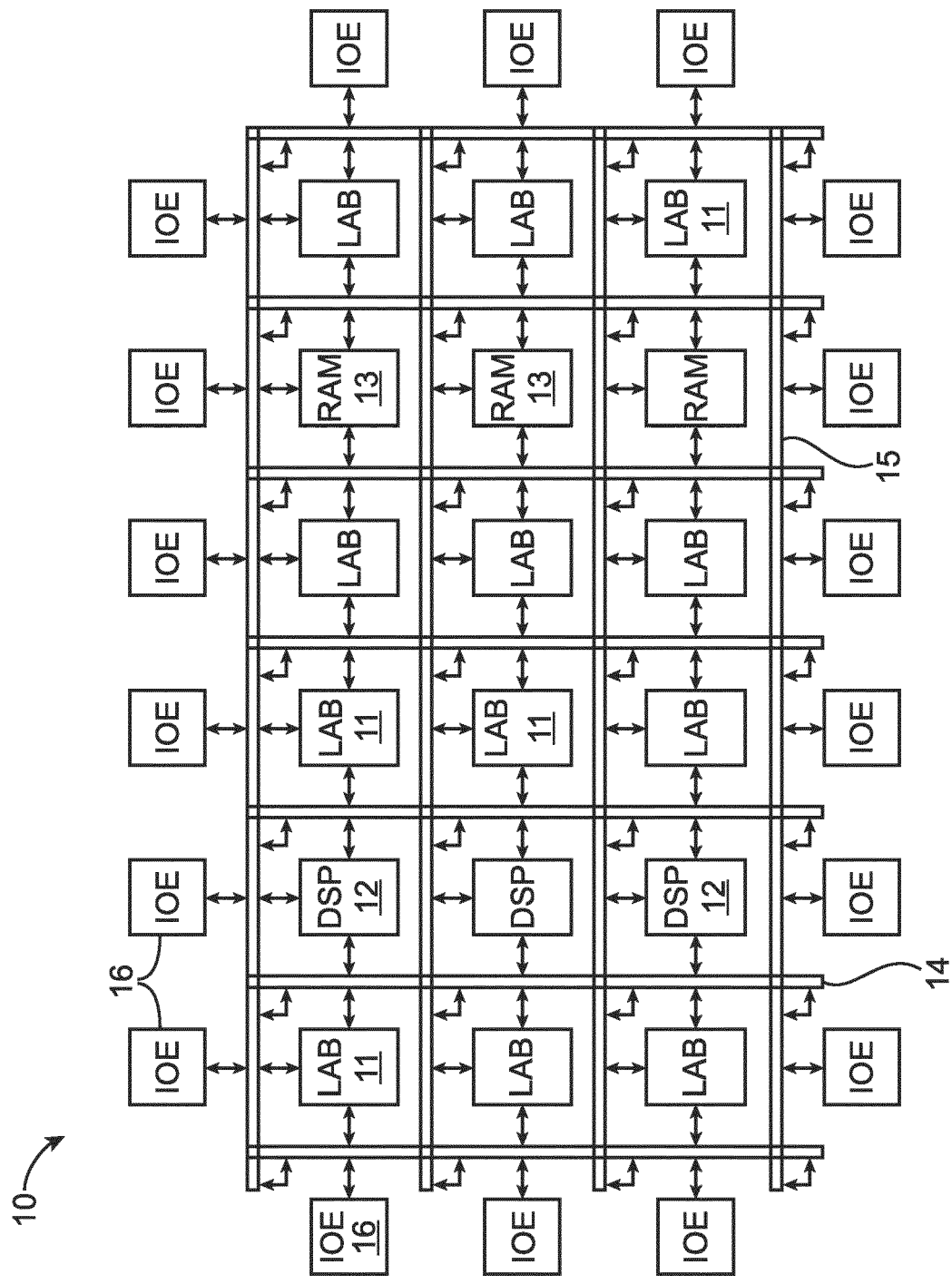


FIG. 6

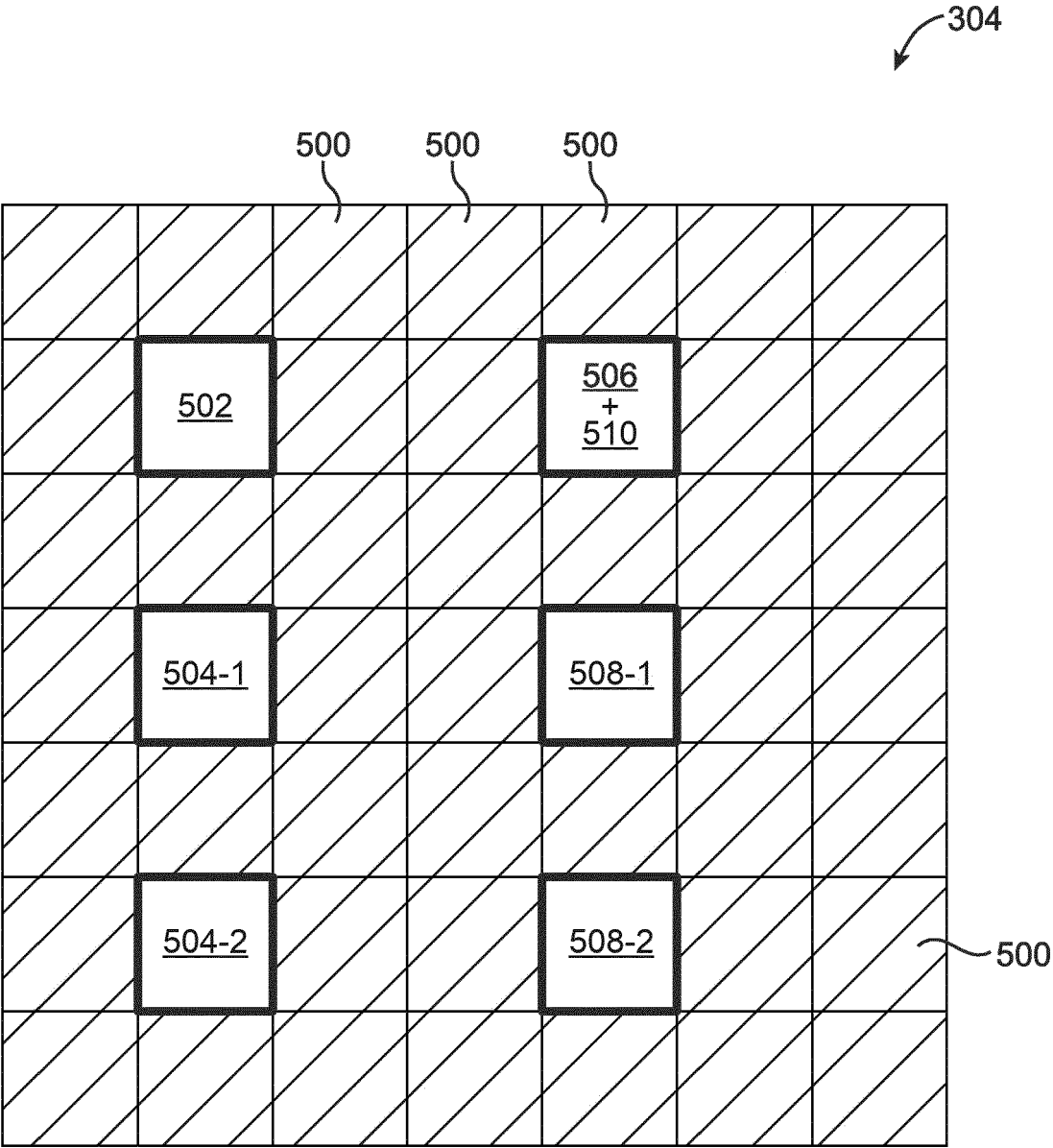


FIG. 7

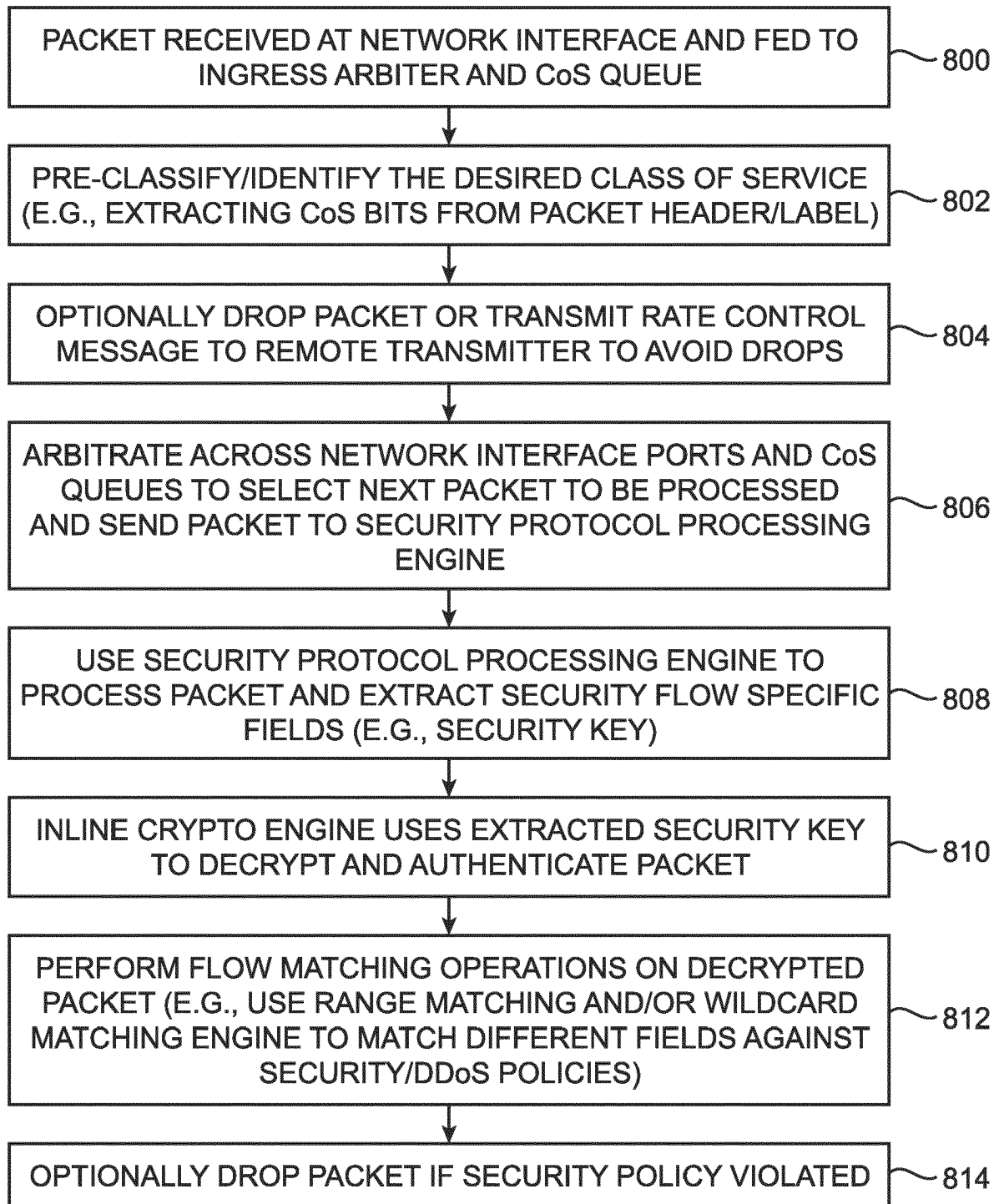


FIG. 8A

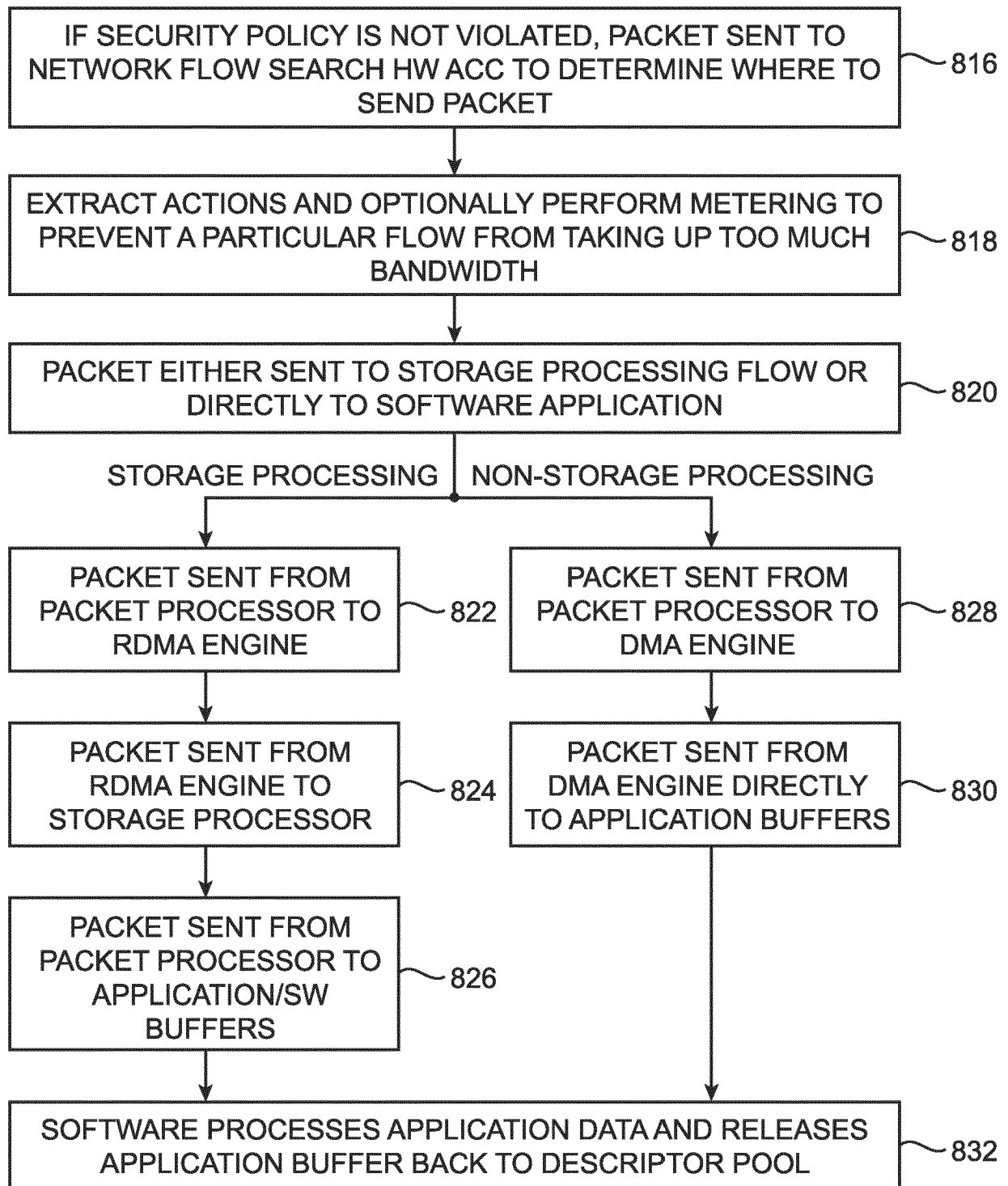


FIG. 8B

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 20190230049 A1 [0002]